DATA SELECTION FOR NOISE ROBUST EXEMPLAR MATCHING

Emre Yılmaz, Jort F. Gemmeke, Hugo Van hamme

Dept. ESAT-PSI, KU Leuven, Belgium

{emre.yilmaz,hugo.vanhamme}@esat.kuleuven.be, jgemmeke@amadana.nl

ABSTRACT

Exemplar-based acoustic modeling is based on labeled training segments that are compared with the unseen test utterances with respect to a dissimilarity measure. Using a larger number of accurately labeled exemplars provides better generalization thus improved recognition performance which comes with increased computation and memory requirements. We have recently developed a noise robust exemplar matching-based automatic speech recognition system which uses a large number of undercomplete dictionaries containing speech exemplars of the same length and label to recognize noisy speech. In this work, we investigate several speech exemplar selection techniques proposed for undercomplete speech dictionaries to find a trade-off between the recognition accuracy and the acoustic model size in terms of the amount of speech exemplars used for recognition. The exemplar selection criterion has be to chosen carefully as the amount of redundancy in these dictionaries is very limited compared to overcomplete dictionaries containing plenty of exemplars. The recognition accuracies obtained on the small vocabulary track of the 2nd CHiME Challenge and the AURORA-2 database using the complete and pruned dictionaries are compared to investigate the performance of each selection criterion.

Index Terms— Noise robust exemplar matching, alpha-beta divergence, collinearity reduction, k-medoids, exemplar selection

1. INTRODUCTION

Exemplar-based speech recognition systems [1–7] use labeled segments from training data to identify unseen speech. These approaches resemble the first attempts to solve the automatic speech recognition (ASR) problem performing dynamic time warping [8–10]. The recognition can be performed by comparing these labeled segments with the segments from the test utterances with respect to a dissimilarity measure. Though exemplars provide the most natural duration and trajectory modeling when compared to its statistical counterparts, e.g. hidden Markov models (HMM) or deep neural networks (DNN), large amounts of data are required to handle the acoustic variation among different utterances.

In order to reduce high memory and computational power requirements, several exemplar selection algorithms are proposed in [11, 12]. The main goal of these techniques is to remove less informative exemplars, e.g. duplicates or rarely used ones, or whose presence result in inaccurate recognition and achieve comparable recognition accuracies using only a portion of the exemplars. Statistical acoustic model training also benefits from data selection as the training times are reduced significantly and sometimes the recognition performance is improved due to the reduced noise and redundancy in the training data [13–16].

This research was funded by the KU Leuven research grant GOA/14/005 (CAMETRON). Emre Yılmaz is now with Radboud University of Nijmegen, Netherlands. Jort F. Gemmeke is now with Google Inc., USA.

Using exemplars in a sparse representation (SR) formulation provides significantly improved noise robustness and exemplar-based sparse representations have been successfully used for feature extraction, speech enhancement and noise robust speech recognition tasks [17–20]. These approaches model the acoustics using fixed length exemplars which are labeled at frame level and stored in the columns of a single overcomplete dictionary. Noisy speech segments are jointly approximated as a sparse linear combination of speech and noise exemplars with exemplar weights obtained by solving a regularized convex optimization problem.

Reducing the dimensions of large datasets stored in a single overcomplete dictionary has been investigated in different fields and several matrix decompositions such as the singular value decomposition (SVD), rank revealing QR decomposition, CUR matrix decomposition, interpolative decomposition (ID) have been used to obtain a low-rank matrix approximation of the complete data matrix [21]. Although the SVD is known to provide the best rank-k approximation, interpretation of the principal components is difficult in data analysis [22]. Therefore, several CUR matrix decompositions have been proposed in which a matrix is decomposed as a product of three matrices $\mathbf{C}$, $\mathbf{U}$, $\mathbf{R}$ and the matrices $\mathbf{C}$ and $\mathbf{R}$ consist of a subset of the actual columns and rows respectively [23, 24]. Several computationally efficient exemplar selection techniques are introduced and applied to polyphonic music transcription task using an overcomplete dictionary containing exemplars of different musical notes in [25]. [26] discusses various ways of reducing the speech and noise dictionaries for an exemplar-based sparse representations approach applied on noise robust ASR task.

In this paper, we focus on the noise robust exemplar matching (N-REM) framework [27] which is an exemplar matching recognition system with noise modeling capabilities. In this framework, the recognizer uses different length exemplars organized in separate dictionaries based on their duration and label (the associated speech unit) [27]. The input speech segments are approximated in a sparse representations formulation, i.e. as a linear combination of the exemplars in each dictionary. Compared to a system using fixed-length exemplars stored in a single dictionary, using separate dictionaries for each class provides better classification as input speech segments are approximated as a combination of exemplars belonging to the same class only. Moreover, each exemplar is associated with a single speech unit and the natural duration distribution of each speech unit in the training data is preserved yielding exemplars of different lengths. This recognizer adopts a reconstruction error based backend, i.e. the recognition is performed by comparing the approximation quality for different classes quantified by a divergence measure and choosing the class sequence that minimizes the total reconstruction error. In [28], we have proposed to use the alpha-beta divergence [29] in place of the generalized Kullback-Leibler divergence (KLD) which has been shown to be more robust against background noise.

The exemplar selection techniques discussed in this paper differ from previous work as the dictionaries store a lot less exemplars due to the use of multiple dictionaries for each exemplar length and label. Compared to the overcomplete dictionaries with a large number of data points, the redundancy in the undercomplete dictionaries used by N-REM is quite limited. Therefore, removing a few informative data points may already result in significant decreases in the recognition accuracy. We have presented the initial findings of our efforts to select a subset of speech exemplars in [30] and reported some promising recognition results on a clean digit recognition task. In this work, we extend the investigation of the proposed exemplar selection technique with the best performance, namely *collinearity reduction*, on all available SNR levels of the small vocabulary track of the 2nd CHiME Challenge and the AURORA-2 database. Moreover, in addition to this technique, we propose a symmetric AB-divergence-based k-medoids algorithm for exemplar selection from undercomplete dictionaries. The AB-divergence is chosen as a dissimilarity measure to be consistent with the recognition setup.

2. NOISE ROBUST EXEMPLAR MATCHING

Training frame sequences representing various speech units (speech exemplars) are extracted based on the state-level alignments obtained using a conventional HMM-based recognizer. Speech exemplars, each comprised of D mel frequency bands and spanning l frames, are reshaped into a single vector and stored in the columns of a speech dictionary $\mathbf{S}_{c,l}$: one for each class c and each frame length l. Each dictionary is of dimensionality $Dl \times N_{c,l}$ where $N_{c,l}$ is the number of available speech exemplars of class c and frame length l. Similarly, a noise dictionary $\mathbf{N}_l$ for each frame length l is formed by reshaping the noise exemplars. Each speech dictionary is concatenated with the noise dictionary of the same length to form a combined dictionary $\mathbf{A}_{c,l} = [\mathbf{S}_{c,l}\, \mathbf{N}_l]$ of dimensionality $Dl \times M_{c,l}$ where $M_{c,l}$ is the total number of available speech and noise exemplars.

An observed noisy (and/or reverberated) speech segment of frame length T frames is also reshaped into vectors by applying a sliding window approach [18] with window length of l frames and stored in an observation matrix $\mathbf{Y}_l = [\mathbf{y}_l^1, \mathbf{y}_l^2 ..., \mathbf{y}_l^{(T-l+1)}]$ of dimensionality $Dl \times (T - l + 1)$. Due to multiple-length exemplars, the window length l is varied between the minimum exemplar length l_{min} and maximum exemplar length l_{max} yielding observation matrices $\mathbf{Y}_l$ for $l_{\text{min}} \leq l \leq l_{\text{max}}$. For every class c, each observation vector $\mathbf{y}_l$ is expressed as a linear combination of the exemplars that are stored in the dictionaries of the same length: $\mathbf{y}_l \approx \sum_{m=1}^{M_{c,l}} x_{c,l}^m \mathbf{a}_{c,l}^m = \mathbf{A}_{c,l}\mathbf{x}_{c,l}$ for $x_{c,l}^m \geq 0$. Here, $\mathbf{x}_{c,l}$ is an $M_{c,l}$-dimensional non-negative weight vector. The combined dictionaries consisting of speech and noise exemplars are presumed to model all acoustic variability in the observed signal due to pronunciation variation, background noise and so forth. This model can also model reverberation by storing reverberated speech exemplars rather than clean speech exemplars.

The exemplar weights $\mathbf{x}_{c,l}$ are obtained by minimizing the cost function consisting of a single term which quantifies the approximation error $d(\mathbf{y}_l, \mathbf{A}_{c,l}\mathbf{x}_{c,l})$ for non-negative exemplar weights. This optimization problem can be solved with non-negative sparse coding (NSC) [31]. The value of approximation error is highly dependent on the divergence measure d and the representation of speech and noise sources. Motivated by its capabilities to weight and scale the individual ratios of the noisy speech and its approximation, $\mathbf{y}_l^i / \hat{\mathbf{y}}_{c,l}^i$ where $\hat{\mathbf{y}}_{c,l} = \mathbf{A}_{c,l}\mathbf{x}_{c,l}$, the AB divergence is used for d. The AB divergence $d_{AB}^{(\alpha,\beta)}(\mathbf{y}, \hat{\mathbf{y}})$ is defined as

$$
= \begin{cases}
-\frac{1}{\alpha\beta} \sum_{k=1}^{K} \left(y_k^\alpha \hat{y}_k^\beta - \frac{\alpha}{\gamma} y_k^\gamma - \frac{\beta}{\gamma} \hat{y}_k^\gamma \right) & \text{for} \quad \alpha, \beta, \gamma \neq 0, \\
\frac{1}{\alpha^2} \sum_{k=1}^{K} \left(y_k^\alpha \log(\frac{y_k^\alpha}{\hat{y}_k^\alpha}) - y_k^\alpha + \hat{y}_k^\alpha \right) & \text{for} \quad \alpha \neq 0, \beta = 0 \\
\frac{1}{\alpha^2} \sum_{k=1}^{K} \left(\log(\frac{\hat{y}_k^\alpha}{y_k^\alpha}) + \frac{y_k^\alpha}{\hat{y}_k^\alpha} - 1 \right) & \text{for} \quad \alpha = -\beta \neq 0, \\
\frac{1}{\beta^2} \sum_{k=1}^{K} \left(\hat{y}_k^\beta \log(\frac{\hat{y}_k^\beta}{y_k^\beta}) - \hat{y}_k^\beta + y_k^\beta \right) & \text{for} \quad \alpha = 0, \beta \neq 0, \\
\frac{1}{2} \sum_{k=1}^{K} \left(\log(y_k) - \log(\hat{y}_k) \right)^2 & \text{for} \quad \alpha, \beta = 0
\end{cases}
\tag{1}
$$

where $\gamma = \alpha + \beta$. The two parameters of the AB divergence can be automatically adjusted based on the amount of contamination in the target utterance as the recognition performance for different noise levels depends on the emphasized (reliable) time-frequency bins. For the NSC solution, we apply the multiplicative update rule minimizing the approximation error $d(\mathbf{y}_l, \mathbf{A}_{c,l}\mathbf{x}_{c,l})$ using the AB divergence for $\alpha \neq 0$ which is given by

$$
\mathbf{x}_{c,l} \leftarrow \left(\mathbf{x}_{c,l} \odot \left((\mathbf{A}_{c,l}^T \mathbf{Z}_{c,l}) \oslash (\mathbf{A}_{c,l}^T (\mathbf{A}_{c,l}\mathbf{x}_{c,l})^{\cdot[\gamma-1]}) \right)^{\cdot[\omega/\alpha]} \right)^{\cdot[1+\theta]},
\tag{2}
$$

where $\mathbf{Z}_{c,l} = \mathbf{y}_l^{\cdot[\alpha]} \odot (\mathbf{A}_{c,l}\mathbf{x}_{c,l})^{\cdot[\beta-1]}$ and $^{\cdot[\,]}$ denotes element-wise exponentiation. ω is a value between $(0, 2)$ and θ is a very small positive number [32].

All observation matrices Y_l for $l_{\text{min}} \leq l \leq l_{\text{max}}$ are approximated using the combined dictionaries $\mathbf{A}_{c,l}$ of the corresponding length by applying the multiplicative update rule. To quantify the approximation quality, we use the reconstruction error between the noisy speech segments and their approximations. The multiplicative update rule is applied iteratively until the reconstruction error provides enough discrimination between different classes. The number of iterations that satisfies this criterion has been investigated in pilot experiments. After a fixed number of iterations for all dictionaries, the reconstruction errors between the observation matrix $\mathbf{Y}_l$ and its approximations $\mathbf{A}_{c,l}\mathbf{x}_{c,l}$ are calculated for $l_{\text{min}} \leq l \leq l_{\text{max}}$. As the label of each dictionary is known, decoding is performed by applying dynamic programming [33] to find the class sequence that minimizes the reconstruction error (taking the grammar into account if necessary).

3. EXEMPLAR SELECTION TECHNIQUES

The N-REM recognition scheme benefits from discarding redundant speech exemplars due to two main reasons. First, the computational load mainly due to the iterative evaluation of the multiplicative update rule reduces proportional to the dictionary sizes. Furthermore, the memory required to store the pruned dictionaries is much less than storing the complete dictionaries. For this purpose, we investigate the impact of two exemplar selection methods, namely collinearity reduction and k-medoids with symmetric AB divergence, on the recognition accuracy in both clean and noisy conditions.

3.1. Collinearity Reduction (CR)

The CR selection technique discards exemplars that are well approximated by the other exemplars of the same length and class (i.e. other exemplars in the same dictionary). The exemplars with larger reconstruction errors are expected to contribute more when approximating unseen noisy segments compared to the ones with smaller reconstruction errors. Therefore, the CR technique compares the reconstruction errors for all exemplars in a dictionary by approximating each exemplar as a linear combination of the other exemplars in the same dictionary. This idea is applied iteratively by removing the exemplar that is approximated with the minimum reconstruction error

at each iteration until the minimum number of exemplars requirement in a dictionary is met.

3.2. K-medoids with AB Divergence (KMED)

KMED selection technique is based on the partitioning around medoids (PAM) technique [34] using a symmetric version of the AB divergence as a novel dissimilarity measure. The symmetric version of the AB divergence given in Equation (1) is obtained as $\frac{1}{2}\left[d_{AB}^{(\alpha,\beta)}(\mathbf{y},\hat{\mathbf{y}}) + d_{AB}^{(\alpha,\beta)}(\hat{\mathbf{y}},\mathbf{y})\right]$. The higher computational complexity of the PAM technique mentioned in [35] is not valid in this scenario as the number of speech exemplars in each dictionary is mostly on the order of magnitude one and two. This selection technique is applied to every dictionary to obtain a certain number of medoids that are expected to represent the convex hull formed by the complete dictionary accurate enough. The divergence parameters are chosen based on the recognition performance of the speech dictionaries on clean speech and the ones providing the best clean speech recognition performance are used during the exemplar selection.

4. EXPERIMENTAL SETUP

4.1. Databases

The training material of AURORA-2 [36] consists of a clean and a multi-condition training set, each containing 8440 utterances. The multi-condition training set was constructed by mixing the clean utterances with noise at SNR levels of 20, 15, 10 and 5 dB. Test set A and B consists of 4 clean and 24 noisy datasets at six SNR levels between -5 and 20 dB. The noise types of test set A match the multi-condition training set. Each subset contains 1001 utterances with one to seven digits 0-9 or oh. To reduce the simulation times, we subsampled the test sets by a factor of 4 (1000 utterances per SNR).

The small vocabulary track of the 2$^{\text{nd}}$ CHiME Challenge [37] addresses the problem of recognizing commands in a noisy and reverberant living room. The clean utterances contain utterances from 34 speakers reading 6-word sequences of the form *command-color-preposition-letter-digit-adverb*. There are 25 different letters, 10 different digits and 4 different alternatives for each of the other classes. The recognition accuracy of a system is calculated based on the correctly recognized letter and digit keywords.

4.2. Dictionary Creation and Implementation Details

The speech exemplars of AURORA-2 data are extracted from the clean training set. Acoustic feature vectors are represented in mel-scaled magnitude spectra with 23 frequency bands. The speech exemplars representing half-digits are segmented by a conventional HMM-based system. The complete dictionary contains in total 52,295 speech exemplars excluding 990 silence exemplars. The number of noise exemplars varies depending on the duration of the noise-only sequences that are selected by active noise exemplar selection (ANES) [27]. On average, the recognizer with the pruned dictionaries containing 20% of the exemplars in each dictionary uses 11,355 and 1,044 noise exemplars/utterance in total at SNR level of -5 dB and clean speech respectively. The divergence parameters (α,β) for the KMED selection technique are set to 1 and 0.25 respectively. The minimum and maximum exemplar lengths are 8 and 40 frames respectively. Exemplars longer than 40 frames are omitted to limit the number of dictionaries. The recognizer uses 675 class-dependent dictionaries in total. In the recognition phase, noise dictionaries are created by performing noise sniffing and ANES. The combined dictionaries and observation matrices are l_2-normalized for all SNR levels. The multiplicative update rule is iterated 100 times for convergence of all frame lengths. The further details are given in [28]. The word error rate (WER) has been used to quantify the recognition accuracy for the AURORA-2 digit recognition task.

The exemplars and noisy speech segments from CHIME-2 data are represented as mel-scaled magnitude spectral features extracted with a 26 channel mel-scaled filter bank ($D = 26$). The frame length is 25 ms and the frame shift is 10 ms. The binaural data is averaged in the spectral domain to obtain 26-dimensional feature vectors. The exemplars are extracted from the reverberated utterances in the training set according to the state-based segmentations obtained using the acoustic models in the toolkit provided with the database. Exemplars belonging to each speaker are organized in separate dictionary sets for speaker-dependent modeling yielding 34 different dictionary sets. Based on the availability of the exemplars, the minimum and maximum exemplar lengths are 4 and 40 frames respectively. Half-word exemplars seemed to generalize sufficiently to unseen data for the recognition task. Dictionary sizes vary with different classes and speakers. The divergence parameters (α,β) for the KMED selection technique are set to 1 and 0 respectively. *Prewarping* [38] is applied to boost the modeling capabilities of the underpopulated speech dictionaries (especially for the ones belonging to letters due to the high number of alternatives and hence the small number of exemplars per class) and it is limited to a single frame. The number of exemplars in each dictionary after prewarping is limited to 50. The noise modeling is detailed in [28]. The multiplicative update rule is iterated 25 times to obtain the exemplar weights. The columns of the combined dictionaries and observation matrices are l_2-normalized. The further details are given in [28]. The keyword recognition accuracy (RA) is used to evaluate the system performance on the CHIME-2 data.

5. RESULTS AND DISCUSSION

The exemplar selection techniques described in Section 3 are applied to the speech dictionaries obtained from AURORA-2 and CHIME-2 data and the recognition performance of the recognizers using only 20% of the exemplars per dictionary are presented in Table 1 and 2. The results obtained using the N-REM recognizer using the generalized KLD (N-REM-KLD) with the complete dictionaries [28], multi-condition trained GMM-HMM and DNN-HMM recognizers and other exemplar-based sparse representation systems [18], namely sparse classification (SC) and feature enhancement (FE), are also provided for comparison. The baseline results obtained with the complete dictionaries and the best results provided by the pruned dictionaries are given in bold.

A pruning rate of 80%, i.e. using 20% of the exemplars in a dictionary, is chosen based on the initial results presented in [30]. This choice aims to compare the amount of degradation in the recognition accuracy when pruning goes further than the *safe* pruning rate of 70% which is defined as the largest pruning rate without significant recognition accuracy loss [30]. We compare CR and KMED techniques with the CUR decomposition which is a randomized column selection algorithm proposed as a part of the CUR matrix decomposition in [22]. This algorithm randomly selects a subset of the columns of a data matrix with respect to the probability distribution computed as the normalized statistical leverage scores. The CUR decomposition has been successfully applied in selecting a very small number of exemplars from an overcomplete dictionary without a significant recognition accuracy loss. We further provide the recognition accuracies obtained using the randomly pruned dictionaries (RND).

The WERs obtained on the clean test set of AURORA-2 are presented in the middle panel of Table 1a and 1b. The N-REM performance using the complete dictionaries is given in the first row of the tables. The clean speech performance of CR is the best among the

Table 1: Word error rates in % obtained on test set A and B of AURORA-2 using 20% of exemplars in each dictionary

(a) Test set A

SNR(dB)	clean	-5	0	5	10	15	20	*0-20*
N-REM	**1.8**	**14.9**	**8.5**	**5.8**	**4.7**	**3.5**	**2.3**	**5.0**
CR	**2.8**	19.8	**10.8**	8.0	**6.3**	**4.7**	**3.5**	**6.7**
KMED	3.0	**18.6**	10.9	**7.9**	6.4	5.0	4.1	6.9
CUR	4.1	20.2	12.6	9.0	7.4	5.6	4.5	7.8
RND	4.1	20.4	12.5	9.0	7.2	5.5	4.5	7.8
N-REM-KLD	1.7	19.1	9.2	5.9	4.9	3.6	2.4	5.2
DNN	0.5	52.4	17.9	3.8	1.5	0.9	0.7	5.0
GMM	0.7	60.8	24.3	7.3	2.9	1.3	0.8	7.3
SC	3.7	35.2	13.8	7.4	5.6	4.8	4.5	7.2
FE	0.5	30.4	10.7	3.3	1.5	1.1	0.7	3.5

(b) Test set B

SNR(dB)	clean	-5	0	5	10	15	20	*0-20*
N-REM	**1.8**	**53.5**	**24.5**	**10.4**	**4.9**	**3.1**	**2.5**	**9.0**
CR	**2.8**	56.7	27.5	12.5	7.0	**4.7**	**3.5**	11.0
KMED	3.0	58.5	**25.9**	**11.7**	**6.9**	5.0	4.6	**10.8**
CUR	4.1	57.6	26.4	13.0	7.4	5.7	5.1	11.5
RND	4.1	**56.1**	26.9	12.8	7.2	5.6	4.3	11.4
N-REM-KLD	1.7	55.0	24.3	10.1	5.5	3.5	2.7	9.2
DNN	0.5	62.9	24.3	6.9	2.0	1.1	0.5	6.7
GMM	0.7	64.0	25.9	7.4	2.6	1.2	0.9	7.6
SC	3.7	52.4	23.5	11.0	5.9	2.7	4.5	9.9
FE	0.5	52.6	20.5	5.7	2.1	1.2	0.5	6.0

Table 2: Keyword recognition accuracies in % obtained on the dev. and test set of CHIME-2 using 20% of exemplars in each dictionary

(a) Development Set

SNR(dB)	-6	-3	0	3	6	9	*Avg*
N-REM	**75.4**	**78.8**	**86.3**	**90.5**	**91.2**	**92.7**	**85.8**
CR	71.5	77.7	83.6	90.0	90.6	92.3	84.3
KMED	**73.0**	**77.8**	**84.7**	**90.3**	**91.3**	**92.4**	**84.9**
CUR	69.3	76.3	82.3	87.9	89.7	91.9	82.9
RND	70.4	76.1	81.8	88.8	89.2	91.5	83.0
N-REM-KLD	70.4	77.9	84.8	90.4	92.6	93.8	85.0
GMM	49.3	58.7	67.5	75.1	78.8	82.9	68.7
FE	68.0	72.2	80.9	86.7	89.0	90.5	81.2
HMM-FE	69.1	73.6	81.5	87.3	89.4	90.3	81.9
SC	75.5	81.4	87.5	89.9	92.4	92.3	86.5

(b) Test set

SNR(dB)	-6	-3	0	3	6	9	*Avg*
N-REM	**73.9**	**79.7**	**86.1**	**88.0**	**90.9**	**92.6**	**85.2**
CR	**72.1**	**78.7**	**84.9**	**87.1**	**90.6**	**91.8**	**84.2**
KMED	71.8	77.9	83.8	86.9	89.4	91.6	83.6
CUR	70.1	77.4	82.9	85.5	88.7	90.4	82.5
RND	70.6	77.3	82.9	86.0	88.6	90.5	82.7
N-REM-KLD	71.0	78.9	85.3	88.7	91.9	92.8	84.8
HMM	49.7	57.9	67.8	73.7	80.8	82.7	68.8
FE	67.2	75.9	81.1	86.4	90.7	92.0	82.2
HMM-FE	67.0	77.0	81.8	87.0	91.2	92.4	82.7
SC	76.5	81.3	88.9	90.5	92.7	93.2	87.2

results obtained with the pruned dictionaries with a WER of 2.8% compared to 1.8% yielded by the complete dictionaries. KMED also provides a comparable result with a WER of 3.0%. These results are consistent with the clean speech recognition results of CR presented in [30]. Dictionaries pruned with the other techniques yield worse performance.

The results on the noisy sets of test set A are given in the rightmost panel of Table 1a. These results further demonstrate the effectiveness of CR and KMED in the noisy scenarios. N-REM with complete dictionaries has a WER of 5.0% on average. CR and KMED provide a WER of 6.7% and 6.9% respectively. CUR performs as poorly as RND on this exemplar selection task yielding a WER of 7.8%. The results on test set B, which are presented in Figure 1b, show a similar trend and the best results in the mismatched noise case are obtained using the dictionaries pruned by CR at high SNR levels and by KMED at low SNR levels. At -5 dB of test set B, RND provides the best results which is explained by the minor impact of the speech dictionaries on the recognition accuracy due to very poor noise modeling. CR and KMED perform better than CUR and RND on average similar to the matched noise case.

The RAs obtained on the development and test sets of CHIME-2 data are shown in Table 2. On the development set, KMED and CR yield an average RA of 84.9% and 84.3% compared to 85.8% of the N-REM baseline. CUR and RND have a comparable RA of 82.9% and 83.0% respectively. On the test set, CR provides an average RA of 84.2% which is slightly better than 83.6% of KMED. These results are higher than 82.5% of CUR and 82.7% of RND.

From these results, it can be concluded that CR and KMED techniques achieve effective exemplar selection from undercomplete dictionaries by reducing the dictionary sizes significantly without a significant loss in the recognition performance, especially at higher SNR levels. Based on the geometrical interpretation of this exemplar selection task as explained in [30], these techniques pick the exemplars that preserve the convex hulls formed by the speech dictionaries in the positive orthant. As a result, the dictionaries pruned by CR and KMED have a more precise description of each speech unit in the high-dimensional feature space compared to the other techniques and the noisy mixtures can still be separated accurately by picking a few number noise and speech exemplars with much less computational and memory requirements compared to the complete dictionaries.

6. CONCLUSION

This paper investigates the performance of several exemplar selection approaches proposed for picking the most informative exemplars from undercomplete dictionaries which are used in the noise robust exemplar matching framework. We first apply the *collinearity reduction* approach, which has shown superior performance on clean speech in previous work, to noisy speech to explore how robust the pruned dictionaries against background noise. Furthermore, we investigate the performance of a k-medoids exemplar selection approach which uses a novel dissimilarity measure, namely the symmetric alpha-beta divergence, in accordance with the recognizer. The dictionaries pruned by both techniques have performed considerably better than random pruning and the column selection of the CUR decomposition which has provided impressive results on overcomplete dictionaries.

7. REFERENCES

[1] M. De Wachter, K. Demuynck, D. Van Compernolle, and P. Wambacq, "Data driven exemplar based continuous speech recognition," in *Proc. European Conf. on Speech Communication and Technology*, 2003, pp. 1133–1136.

[2] T. Deselaers, G. Heigold, and H. Ney, "Speech recognition with state-based nearest neighbour classifiers," in *Proc. INTERSPEECH*, Antwerp, Belgium, 2007, pp. 2093–2096.

[3] L. Golipour and D. O'Shaughnessy, "Context-independent phoneme recognition using a k-nearest neighbour classification approach," in *Proc. ICASSP*, Apr. 2009, pp. 1341–1344.

[4] S. Sundaram and J. R. Bellegarda, "Latent perceptual mapping with data-driven variable-length acoustic units for template-based speech recognition," in *Proc. ICASSP*, 2012, pp. 4125–4128.

[5] G. Heigold, P. Nguyen, M. Weintraub, and V. Vanhoucke, "Investigations on exemplar-based features for speech recognition towards thousands of hours of unsupervised, noisy data," in *Proc. ICASSP*, Kyoto, Japan, 2012, pp. 4437–4440.

[6] T. N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, D. Van Compernolle, K. Demuynck, J. F. Gemmeke, J. R. Bellegarda, and S. Sundaram, "Exemplar-based processing for speech recognition: An overview," *IEEE Signal Processing Magazine*, vol. 29(6), pp. 98–113, Nov. 2012.

[7] X. Sun and Y. Zhao, "Integrated exemplar-based template matching and statistical modeling for continuous speech recognition," *EURASIP Journal on ASMP*, vol. 2014, no. 1, 2014.

[8] H. Sakoe and S. Chiba, "A dynamic programming approach to continuous speech recognition," in *Proceedings of the Seventh International Congress on Acoustics*, 1971, vol. 3, pp. 65–69.

[9] G.M. White and R. Neely, "Speech recognition experiments with linear predication, bandpass filtering, and dynamic programming," *IEEE TASSP*, vol. 24(2), pp. 183–188, Apr. 1976.

[10] L. Rabiner, A.E. Rosenberg, and S.E. Levinson, "Considerations in dynamic time warping algorithms for discrete word recognition," *IEEE TASSP*, vol. 26(6), pp. 575–582, Dec. 1978.

[11] D. Seppi and D. Van Compernolle, "Data pruning for template-based automatic speech recognition," in *Proc. Interspeech*, Sept. 2010, pp. 985–988.

[12] X. Sun and Y. Zhao, "New methods for template selection and compression in continuous speech recognition," in *Proc. Interspeech*, Florence, Italy, Aug. 2011, pp. 985–988.

[13] T. M. Kamm and G. L. Meyer, "Selective sampling of training data for speech recognition," in *Proc. HLT*, 2002, pp. 20–24.

[14] D. Hakkani-Tur, G. Riccardi, and A. Gorin, "Active learning for automatic speech recognition," in *Proc. ICASSP*, May 2002, vol. 4, pp. IV–3904–IV–3907.

[15] A. Nagroski, L. Boves, and H. Steeneken, "In search of optimal data selection for training of automatic speech recognition systems," in *Proc. ASRU*, Nov. 2003, pp. 67–72.

[16] Y. Wu, R. Zhang, and A. Rudnicky, "Data selection for speech recognition," in *Proc. ASRU*, Dec. 2007, pp. 562–565.

[17] T. N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, and A. Sethy, "Sparse representations features for speech recognition," in *Proc. Interspeech*, Sept. 2010, pp. 2254–2257.

[18] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE TASLP*, vol. 19(7), pp. 2067–2080, 2011.

[19] A. Hurmalainen, J.F. Gemmeke, and T. Virtanen, "Non-negative matrix deconvolution in noise robust speech recognition," in *Proc. ICASSP*, May 2011, pp. 4588–4591.

[20] Q. F. Tan and S. S. Narayanan, "Novel variations of group sparse regularization techniques with applications to noise robust automatic speech recognition," *IEEE TASLP*, vol. 20, no. 4, pp. 1337–1346, May 2012.

[21] N. Halko, P.-G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM Review*, vol. 53(2), pp. 217–288, 2011.

[22] M. W. Mahoney and P. Drineas, "CUR matrix decompositions for improved data analysis," *Proceedings of the National Academy of Sciences*, vol. 106, no. 3, pp. 697–702, 2009.

[23] S. A. Goreinov, E. E. Tyrtyshnikov, and N. L. Zamarashkin, "A theory of pseudoskeleton approximations," *Linear Algebra and its Applications*, vol. 261, no. 13, pp. 1–21, 1997.

[24] A. Frieze, R. Kannan, and S. Vempala, "Fast monte-carlo algorithms for finding low-rank approximations," *J. ACM*, vol. 51, no. 6, pp. 1025–1041, Nov 2004.

[25] I. Ari, U. Simsekli, A. T. Cemgil, and L. Akarun, "Randomized matrix decompositions and exemplar selection in large dictionaries for polyphonic piano transcription," *Journal of New Music Research*, vol. 43(3), pp. 255–265, 2014.

[26] A. Hurmalainen, J. F. Gemmeke, and Virtanen T., "Modelling non-stationary noise with spectral factorisation in automatic speech recognition," *Computer Speech & Language*, vol. 27, no. 3, pp. 763–779, 2012.

[27] E. Yılmaz, J. F. Gemmeke, and H. Van hamme, "Noise robust exemplar matching using sparse representations of speech," *IEEE/ACM TASLP*, vol. 22(8), pp. 1306–1319, Aug. 2014.

[28] E. Yılmaz, J. F. Gemmeke, and H. Van hamme, "Noise robust exemplar matching with alpha-beta divergence," *Accepted for publication, Speech Comm.*, 2015.

[29] A. Cichocki, S. Cruces, and S.-I. Amari, "Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization," *Entropy*, vol. 13, pp. 134–170, 2011.

[30] E. Yılmaz, J. F. Gemmeke, and H. Van hamme, "Exemplar selection techniques for sparse representations of speech using multiple dictionaries," in *Proc. EUSIPCO*, Sept. 2013, pp. 1–5.

[31] P.O. Hoyer, "Non-negative sparse coding," in *IEEE Workshop on Neural Networks for Signal Processing*, 2002, pp. 557–565.

[32] A. Cichocki, R. Zdunek, and S. Amari, "Csiszár's divergences for non-negative matrix factorization: family of new algorithms," in *Proc. Int. Conf. on Independent Component Analysis and Blind Signal Separation*, 2006, pp. 32–39.

[33] H. Ney, "The use of a one-stage dynamic programming algorithm for connected word recognition," *IEEE TASSP*, vol. 32, no. 2, pp. 263–271, Apr 1984.

[34] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, 9th edition, Mar. 1990.

[35] J. Han, M. Kamber, and A. K. H. Tung, "Spatial clustering methods in data mining: A survey," in *Geographic Data Mining and Knowledge Discovery, Res. Monographs in GIS*, 2001.

[36] H. G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ISCA Tutorial and Research Workshop ASR2000*, Sept. 2000, pp. 181–188.

[37] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines," in *Proc. ICASSP*, May 2013, pp. 126–130.

[38] E. Yılmaz, J. F. Gemmeke, and H. Van hamme, "Noise-robust automatic speech recognition with exemplar-based sparse representations using multiple length adaptive dictionaries," in *Proc. CHIME-2013*, June 2013, pp. 39–43.